

An Egocentric Look at Video Photographer Identity

Yedid Hoshen Shmuel Peleg
The Hebrew University of Jerusalem
Jerusalem, Israel

Abstract

Egocentric cameras are being worn by an increasing number of users, among them many security forces worldwide. GoPro cameras already penetrated the mass market, reporting substantial increase in sales every year. As head-worn cameras do not capture the photographer, it may seem that the anonymity of the photographer is preserved even when the video is publicly distributed.

We show that camera motion, as can be computed from the egocentric video, provides unique identity information. The photographer can be reliably recognized from a few seconds of video captured when walking. The proposed method achieves more than 90% recognition accuracy in cases where the random success rate is only 3%.

Applications can include theft prevention by locking the camera when not worn by its rightful owner. Searching video sharing services (e.g. YouTube) for egocentric videos shot by a specific photographer may also become possible. An important message in this paper is that photographers should be aware that sharing egocentric video will compromise their anonymity, even when their face is not visible.

1. Introduction

The popularity of head worn egocentric cameras is increasing. GoPro reports an increase in sales of 66% every year, and cameras are widely used by extreme sports enthusiasts and by law enforcement and military personnel.

Special features of egocentric video include:

- The camera is worn by the photographer, and is recording while the photographer performs normal activities.
- The camera moves with the photographer's head.
- The camera does not record images of the photographer. In spite of this we show that photographers can often be identified.

Photographers feel secure that sharing their egocentric videos on social media does not compromise their identity (Fig. 1). Police forces routinely release footage of officer



Figure 1. a) A GoPro video uploaded to YouTube allegedly capturing a crime from the POV of the robber. Can the robber be recognized? b) A GoPro video uploaded by US soldiers in combat. Are their identities safe?

activity and operations of special forces recorded by wearable cameras are widely published on YouTube. Some have even recorded and published their own crimes. A consequence of our work is that the photographer identity of such videos can sometimes be found from camera motion.

Body motion is an accurate and replicable feature for identifying people over time. It is often recorded by accelerometers ([16]) or by an overlooking camera. Egocentric video can effectively serve as a head mounted visual gyroscope and can accurately capture body motion information. It follows that any egocentric video which includes walking contains body motion information that can accurately reveal the photographer.

Specifically, we use sparse optical flow vectors (50 flow vectors per frame) taken over a few steps (4 seconds). This results in a set of time-series, one for each component of each optical flow vector. In Fig 2 we show the temporal Fourier Transform of one flow vector for three different sequences, showing visible differences between different photographers.

As a first approach for determining photographer identity, we computed LPC (Linear Predictive Coding ¹) [3] coefficients for each of the optical flow time series. All LPC coefficients of all optical flow sequences were used as a descriptor. Photographer recognition using a non-linear SVM trained on the LPC descriptor gave 81% identification accu-

¹The LPC coefficients of a time series are k values that when scalar multiplied with the last k measurements of the time series, will optimally predict the next measurement.

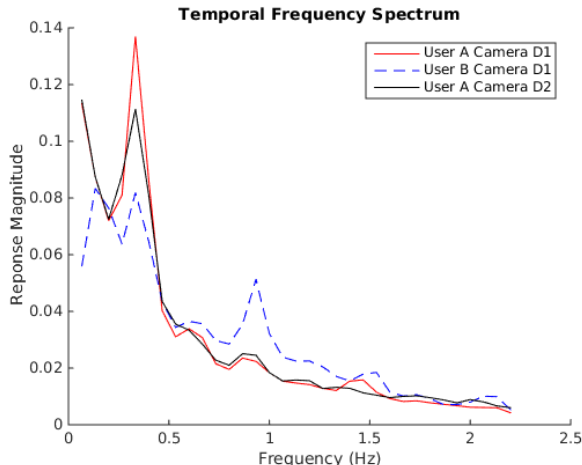


Figure 2. Comparison of the temporal frequency spectra for three videos. Two videos were recorded using camera D1 by users A and B, the third video was recorded by user A using camera D2. It is readily seen that the spectra of the two videos recorded by photographer A are very similar to each other despite being recorded by different cameras and at different times. This suggests that a photographer’s physique is expressed in the motion observed in his video.

racy (vs. accuracy of 3% in random) and verification EER (Equal Error Rate) of 10%.

Our second approach learns the descriptor and classifiers using a Convolutional Neural Network (CNN) which includes layers corresponding to body motion descriptor extraction and to recognition. The CNN is trained on the optical-flow features described above. Using CNN improves the results over the LPC coefficients, yielding 90% identification rate (vs. accuracy of 3% in random) and verification EER (Equal Error Rate) of 8%.

The above experiments were performed on both a small (6 person) public dataset [6] (originally collected for Egocentric Activity Analysis) and on a new, larger (32 person) dataset collected by us especially for Egocentric Video Photographer Recognition (EVPR).

The ability to recognize the photographer quickly and accurately can be important for camera theft prevention and for forensic analysis (e.g. who committed the crime). Other applications are web search by egocentric video photographer and organization of video collections. Wearing a mask does not reduce recognition rate, of course.

2. Previous Work

Determination of the painter of an artwork for preventing forgery and fake artists has attracted attention for centuries. Computer vision researchers have presented several approaches for automatic artist and style classification mainly utilizing low-level and object cues [10, 1].

Recognizing the unseen photographer of a picture is an interesting related problem. In this setting the photographic style [24] and the location of the photograph [8, 13] can be used as cues for photographer recognition. Both methods are unable to distinguish between photographers using cameras on default settings (such as most wearable cameras) and at the same locations. Another approach is automatic recognition of the photographer’s reflection (e.g. in the subject’s eyes [18]), but this relies on having reflective surfaces in the pictures.

Photographer recognition from wearable camera video is a novel problem. Such video is jittery due to the motion of the photographer’s head and body. Although typically a nuisance, we show that frame jitters can accurately determine photographer identity.

Human body motion was already used for recognition. Gait recognition is typically done by a video camera observing a person’s shape and dynamic walking style. These features are able to recognize a person accurately [17]. In our scenario, however, the photographer is not seen by the camera which is worn on his head. Recognition from accelerometers carried on the user’s body [16] is also reported. Shiraga et al. [23] studied people recognition wearing a backpack with stereo cameras. Rotation and period of motion were computed using 3D geometry, and users were accurately recognized. Unlike all prior art, we are interested in recognizing photographers of videos taken by standard wearable cameras (e.g. as exist on video sharing websites), nearly all of which are monocular, head or chest mounted.

Using optical flow for activity recognition from head-mounted cameras has been done by [11, 19, 22, 14] and others. Papers [20, 25] used head motion to retrieve head-mounted camera users observed in other videos recorded at the *same time*. We, on the other hand, use camera motion to recognize the users of wearable cameras *across time*.

Feature design for time series data has been extensively studied, particularly for speech recognition systems ([21]). Speaker verification is a long standing problem which is related to this work. Linear Predictive Coding (LPC) descriptors are very popular for speaker recognition [7]. We show that an LPC-based descriptor is highly effective also for user recognition from egocentric camera video.

In this paper we also take an end-to-end approach of learning features along with the classifier, instead of hand designing the features. We perform this using convolutional neural networks (CNN). For an overview of deep networks see [2]. Learned features are sometimes better than hand-designed features [12].

3. Photographer Recognition from Optical Flow

Egocentric video suffers from bouncy and unsteady motion caused by photographer head and body motion. Al-

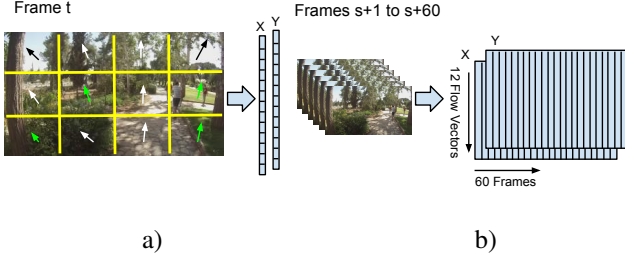


Figure 3. a) 50 Optical flow vectors are calculated for each frame (only 12 shown here), and represented as two columns (each of 50 values), for the x and y optical flow components. b) The feature vector consists of optical flow columns for 60 frames, stacked into two 50×60 arrays, for the x and y components of the flow.

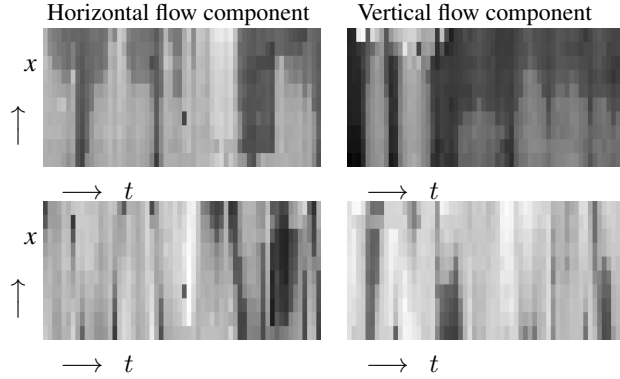


Figure 4. Two examples of the flow feature vectors. Each feature vector consists of 50 optical flow vectors per frame, computed for each of 60 frames. Here only the central row, having 10 flow vectors, is shown. The left and right images show the horizontal and vertical components of the optical flow. Note the rich temporal structure along the time axis.

though usually a nuisance, we show that this motion forms the basis for accurate photographer recognition methods.

We present our basic features in Sec. 3.1. Two alternative descriptors and classifiers are described in Sec. 3.2 and Sec. 3.3.

3.1. Feature Extraction

In the following sections we assume that the video frames were pre-processed in the following way (see Fig. 3):

1. Frames are partitioned into a small number ($m_x \times m_y$) of non-overlapping blocks.
2. $m_x \times m_y$ optical flow vectors are computed for each frame using the Lucas Kanade algorithm [15]. We use 10×5 optical flow vectors per frame.
3. A block of T seconds of such optical flow vectors is taken. We used $T = 4$ seconds, which is long enough to include a few steps. At 15 fps this results in 60 frames.
4. Each feature vector covers a period of 4 seconds, and

we computed feature vectors every 2 seconds. There is an overlap of 2 seconds between two successive feature vectors.

We used optical flow features for photographer recognition, rather than pixel intensities, as the body motion is eventually expressed by the pixel motion. On the other hand, recognition should be invariant to the specific objects seen in the environment, objects that are represented by pixel intensities. CNNs may be able to learn optical flow from pixel intensities, but learning this will require much more data than we can collect.

If dense optical flow were used as a feature, the high feature dimensionality would have lead to overfitting on small datasets. Using a smaller number of flow vectors gave reduced accuracy. In looking for the optimal feature size we found out that a grid size of 10×5 optical flow vectors was a good compromise between overfitting and accuracy.

The feature extraction process is shown in Fig 3. Visualization of two extracted feature vectors is shown in Fig. 4. Full details are in Sec. 6.3.

3.2. LPC Descriptor + Kernel SVM

LPC [3] is a popular time-series descriptor (e.g. for speaker verification). LPC assumes the data is generated by a physical system, here the photographer’s head and body. It attempts to learn a linear regression model for its equations of motion, predicting for each optical flow series the flow value in the next frame given the flow values of previous k frames. Given a feature vector, we calculate an LPC model for each component of each 4s flow time series (100 models in total). Using too few coefficients yields less accurate predictions, while too many coefficients causes overfitting. We found $k=9$ to work well for our case. The final LPC descriptor consisted of all coefficients of all time-series models (100×9).

An RBF-SVM classifier was used for learning both identification (classify LPC descriptor into 1 of M known photographers) and verification (classify LPC descriptor into target photographer or rest-of-the-world). The non-linear (RBF) classifier outperformed linear SVM in almost all cases. As mentioned before, photographer recognition using a non-linear SVM trained on the LPC descriptor gave 81% identification rate (vs. random 3%), and the verification EER (Equal Error Rate) was 10%.

3.3. Convolutional Neural Network

In Sec. 3.2 we described a hand-designed descriptor for identity recognition. The LPC descriptor suffers from several drawbacks:

- The LPC regression model is learned for each time-series separately and ignores the dependence between optical flow vectors.

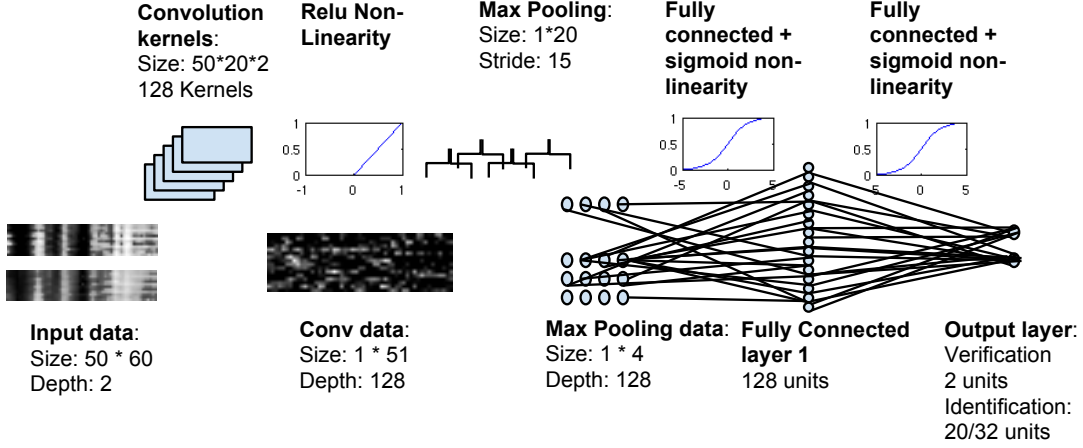


Figure 5. A diagram of our CNN architecture for photographer recognition from a given flow feature vector. The operations on the data are shown on top, the sizes of subsequent data layers are shown on the bottom. The Neural Network learns the descriptor jointly with the classifier, therefore automatically creating a descriptor optimal to this task.

- The LPC descriptor and SVM classifier are learned independently, the labels cannot directly influence the design of the descriptor.

To overcome the above drawbacks, we propose to learn a CNN model for photographer recognition. The CNN learns descriptor and classifier end to end, and is able to take advantage both of dataset labels and the dependence between features when calculating filter coefficients. The CNN is a more general architecture, the LPC descriptor is a subset of descriptors learnable by the network.

Due to the limited number of data points available in our datasets, we limit our CNN to only 2 hidden layers. Using more layers increases model capacity but also increases over-fitting, and this architecture yielded the best performance. The architecture is illustrated in Fig. 5.

Our architecture is tailored especially for egocentric video. As we use sparse optical flow we do not assume much spatial invariance in the features (differently from most image recognition tasks). On the other hand the precise temporal offset of the photographer’s actions is usually not important, e.g. the precise time of the beginning of a photographer’s step is less important than the time between strides. Our architecture should therefore be temporally invariant. The first layer was thus designed to be convolutional in time but not in space.

The kernel size spans all the blocks across the x and y components over K_T frames (we use $K_T = 20$ which is a little longer than the typical step duration). The convolutional layer consists of M kernels (we use $M = 128$). The outputs of the kernels $z_m^1 = W_m * x$ are passed through a ReLU non-linearity ($\max(z_m^1, 0)$). We pool the outputs substantially in time, as the feature vector is of high dimension compared to the amount of training data available. To

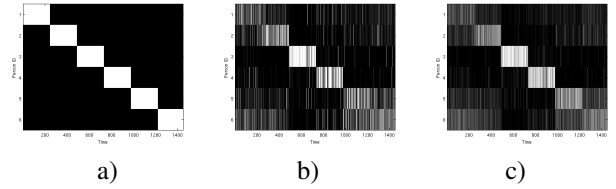


Figure 6. The MAP rule operated on the FPIS dataset: a) Ground truth labels. b) Raw CNN probabilities. c) MAP rule probabilities (for $T=12$ seconds.). The MAP classifier visibly ‘cleaned up’ the prediction.

correspond to the typical time interval between steps we use kernel length of 20 and stride of 15.

The data is then passed through two fully connected (affine) layers each followed by a sigmoid non-linearity ($\sigma(z) = \frac{1}{1+e^{-z}}$). The first fully connected hidden layer has N_1 hidden nodes (we used $N_1 = 128$). The output of this layer is the learned CNN descriptor.

The second fully connected layer is a linear soft-max classifier and has the same number of nodes as the number of output classes: 2 classes for verification, and 20 or 32 classes for identification.

3.4. Joint Prediction from Several Descriptors

Sec. 3.2 and Sec. 3.3 described a method to train a photographer classifier on a short (4 seconds) video sequence. The video used for recognition is usually significantly longer than 4 seconds.

We split the video into 4 second subsequences (overlapping by 2 seconds) and extract their feature vectors V_t (t is the subsequence number). We compute the identity label (L_t) probability distribution for each feature vector V_t using LPC or CNN classifiers trained as de-

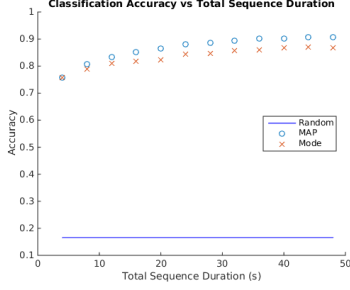


Figure 7. Classification accuracy vs. video length when one feature vector covers $T = 4$ seconds (Using CNN on the FPSI Dataset). Longer video allows extraction of more feature vectors. MAP classification consistently beats mode classification. Both methods can exploit longer sequences and thus improve on 4s sequence recognition. All methods perform far better than random.

scribed above, We then classify the entire video into the globally most likely label, $\text{argmax}_i \prod_t P(L_t = i | V_t) = \text{argmax}_i \sum_t \log(P(L_t = i | V_t))$. While this classifier assumes that feature vectors are IID, we have found that this requirement is not necessary for the success of the method. See Fig. 6 for an example on the FPIS dataset. MAP classification has helped boost the recognition performance on the EVPR dataset to around 90% (an increase of 13%) over the 4s rate.

4. Results

Several experiments were performed to evaluate the effectiveness of our method. As there is no standard dataset for Egocentric Video Photographer Recognition, we use both a small (6 person) public dataset - FPSI [6] that was originally collected for egocentric activity analysis. For each photographer - morning sequences were used for training, and afternoon sequences for testing.

In order to evaluate our method under more principled settings, we collected a new larger (32 person) dataset - EVPR - specifically designed for egocentric photographer video recognition. In the EVPR dataset all photographers recorded two 7 minute sequences (from which we extracted around 200 four second sequences each) on the same day with different head-mounted cameras (D1,D2) for training and testing. 20 of the photographers also recorded another 7 minute sequence with yet another camera (D3) a week later. Both datasets are described in detail in Sec. 6.1. The detailed experimental protocol is described in Sec. 6.2.

4.1. Photographer Identification

Fig. 7 presents the photographer recognition test performance of our network on the FPSI database (6 people). The average correct recognition rate on a single feature vector (describing only 4 seconds of video) is 76% against the random performance of 16.6%.

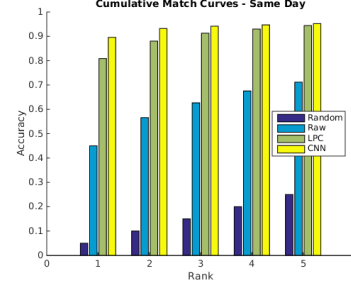


Figure 8. CMC rates for same day recognition (for 12s sequences). LPC accuracy: 81% (Top-1) and 88% (Top-2). The CNN further improves the performance with 90% (Top-1) and 93% (Top-2). Both methods far outperform the random rate of 3% (Top-1) and 6% (Top-2). Both descriptors also beat the raw features by a large margin.

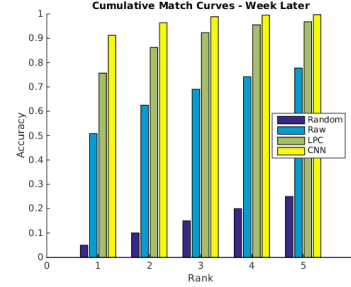


Figure 9. CMC rates for recognition 1 week later (for 12s sequences). LPC accuracy: 76% (Top-1) and 86% (Top-2). The CNN further improves the performance with 91% (Top-1) and 96% (Top-2). Both methods far outperform the random rate of 5% (Top-1) and 10% (Top-2). Both descriptors also beat the raw features by a large margin.

Test videos are usually longer than 4 seconds, and we have multiple feature vectors for each person. We combine predictions over a longer video using the MAP rule in Sec. 3.4. In Fig. 7 we compare the MAP strategy vs. taking the most frequent 4s prediction in the test video (Mode). We observe that using longer sequences further improves recognition performance, reaching around 91% accuracy for 50 seconds of video. We also observe that MAP classifiers consistently beats the Mode classifier and use it in all other experiments.

To evaluate the recognition performance on a larger dataset, we show the performance of our method on our new dataset - EVPR. In this experiment the network was trained on video sequences for each photographer using Camera D1 and is evaluated on video sequences recorded on the same day using Camera D2 and a week later recorded using Camera D3. In Fig. 8 and Fig. 9 we present the cumulative match curve (CMC) for the same day and week later recognition results respectively. We use the Top- k notion, indicating that the correct result appeared within the top k predictions

Descriptor	No Stab		Stab	
	4s	12s	4s	12s
LPC	65%	81%	59%	72%
CNN	77%	90%	71%	86%

Table 1. Same-day CSMC recognition accuracy with and without stabilization.

of the classifier. In addition to LPC and CNN, an RBF-SVM trained on the raw optical flow features is used as baseline to evaluate the quality of our descriptors. High accuracy was achieved in both scenarios, same day CNN recognition accuracy is 90% (top 1) and 93% (top 2). The recognition performance a week later is better with 91% (top 1) and 96% (top 2). The improved performance numbers a week later are expected due to the smaller dataset size (20 vs 32), but are nonetheless encouraging as many photographers were different shoes from the D1 training sequence recorded a week before. This result shows that our method can obtain good recognition performance on meaningful numbers of photographers and across at least a week.

To test the possibility that stabilization would take away some or all the body motion information in the frame motions, the identification experiments were redone with the following pre-processing stage: for each frame (50 flow vectors) the mean framewise vector was calculated and then subtracted from each of the vectors in the frame. As motion between frames is small and some lens distortion correction was performed, this is similar to 2D stabilization. Table 1 shows that such "stabilization" degrades performance somewhat (4-9%), but accuracy still remains fairly high. We note however that more complex stabilization might remove more body motion information.

4.2. Photographer Verification

We also test the verification performance obtained by our method. In order to evaluate verification performance by a single number it is common to use the Equal Error Rate (EER), the error rate at which the False Acceptance Rate (FAR) and False Rejection Rate are equal.

The EER for both the CNN and LPC descriptors for videos of length 4s (one feature vector) and 12s (five feature vectors) is presented in Table 2 while the ROC curves are shown in Fig. 10. A detailed description of our protocol can be found in Sec. 6. It can be seen from our results that high accuracy (low EER) can be obtained by both descriptors: LPC 14% (4s), 10% (12s) and CNN 11% (4s), 8% (12s). The CNN obtains better performance for both durations with a larger improvement for 4s.

It should be noted that all test probe photographers apart from the target photographer had never been used in training. By focusing on modeling the target photographer we can separate him from the rest of the world, and are thus able to generalize to unseen test photographers.

Descriptor	4s	12s
LPC	13.6%	9.6%
CNN	11.3%	8.1%

Table 2. Verification equal error rates for LPC and CNN descriptors with 4s and 12s sequence duration.

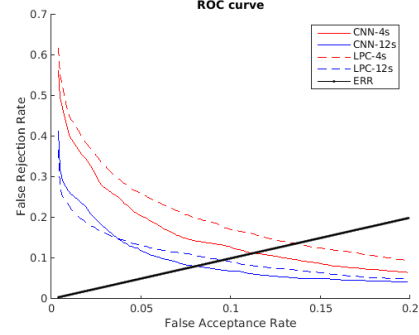


Figure 10. ROC curves for the verification performance of our method for LPC and CNN descriptors of 4s and 12s sequences. For both methods we show the mean ROC curve. The EER of each method is given by the point of intersection between the linear line and its ROC curve.

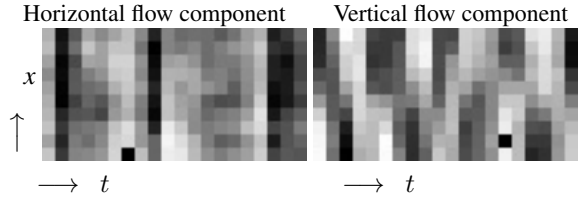


Figure 11. Examples of a temporal filter for the horizontal (left) and vertical (right) flow components. Horizontal axis is time, and vertical axis is location along the central line. The horizontal component filter appears to be sensitive for certain left-right frequencies, while the vertical component filter is sensitive to oscillating rotations: When the right side is moving up the left side is moving down, etc.

5. Discussion

Analysis of CNN features: In order to analyze the features learned by the CNN we visualize the filters learned by the first layer. Fig. 11 shows the horizontal and vertical components of a first layer temporal filter learned by the network. For illustration purposes, only the weights of the central line of pixels are shown. Looking at the weights, we see that the horizontal component filter is tuned to respond to some specific frequencies, while the vertical component looks for sharp rotations. This behavior appears in several other filters suggesting that the network might be using both spectral and transitive cues.

Transfer Learning for verification: In some scenarios it may not be possible to train a verification classifier for each photographer. In such cases Nearest Neighbors may be a good alternative. The following approach is taken: An



Figure 12. Common failure cases for the 4-second descriptor: a-b) Sharp turns of the head result in atypical fast motions, sometimes causing motion blur. c) Large moving objects can also cause atypical optical flow patterns.

identification CNN is trained on half the photographers in the training dataset. We choose a video by a target photographer (that was not used for training the CNN), and extract its CNN descriptors (as in Sec. 3.3), this set of descriptors forms our gallery. Similarly we extract CNN descriptors from all video sequences of photographers not used for training the CNN, this forms our probe set (excluding the sequence used as gallery). For each probe descriptor we check if the euclidean distance from its nearest neighbor in the gallery is smaller than some threshold, and if so we classify it as the target photographer. We used Camera D1 sequences for training and D2 sequences for test. 16 randomly selected photographers were used for training the CNN, and the rest for verification. The same procedure was carried out for LPC (without training a CNN). Multiple 4s sequence predictions are aggregated using simple voting. The average EER for 12s sequences was 15.5% (CNN) and 22%(LPC). Although less accurate than trained classifiers, this shows the network learns to identity features that are general and can be transferred to identify unseen photographers. Nearest Neighbors classification on the optical flow raw features yielded very low performance in accordance with the findings of [20, 25].

Verification on FPSI: We tried learning verification classifiers by choosing one photographer from the FPSI dataset as target, and 4 other photographers as negative training data. The morning sequences of the target photographer were used for training and the afternoon for testing. We tested the classification performance between the afternoon sequences of the target photographer and the remaining 6th non-target photographer from the FPSI dataset. The network however, fit to the train non-target photographers and has not been able to generalize to the unseen probe photographer. We therefore conclude that a significant number of photographers (such as present in the EVPR dataset) is required for training a verification classifier.

Failure cases: In Fig. 12 several cases are shown where the 4 second descriptor failed to give correct recognition. Failure can be caused by sharp head movements (sometimes causing significant blur), by large moving objects, or by lack of features for optical flow computation. It is likely that by identifying such cases and removing their descriptors, higher recognition performance may be achieved.

6. Experimental Procedure

In this section we give a detailed description of the experimental procedure used in Sec. 4.

6.1. Dataset Description

Two datasets were used for evaluation: a public general purpose dataset (FPSI) and a larger dataset (EVPR) collected by us to overcome some of the weaknesses of FPSI.

6.1.1 FPSI Dataset

The First-Person Social Interactions (FPSI) dataset was collected by Fathi et al. [6] for the purpose of activity analysis. 6 individuals (5 males, 1 female) recorded a day’s worth of egocentric video each using head-worn GoPro cameras. Due to battery and memory limitations of the camera, the photographers occasionally took the cameras off and put them on again, ensuring that camera extrinsic parameters were not kept constant.

In this work we learn to recognize video photographers while walking, rather than sitting or standing. We therefore extracted the walking portions of each video using manual labels. It is possible to use a classifier such as described in [19] to find the walking intervals.

6.1.2 EVPR Dataset

The FPSI dataset suffers from several drawbacks: it contains video only for a small number of photographers (6) and each photographer wears the same hat and camera all the time. It is therefore conceivable that learning camera parameters can help recognition. To overcome these issues we collected a larger dataset - Egocentric Video Photographer Recognition (EVPR).

The EVPR consists of head-mounted video sequences collected from 32 photographers. Each video sequence was recorded with a GoPro camera attached to a baseball cap worn on the photographer’s head (as in Fig. 13). Each photographer was asked to walk normally for around 7 minutes along the same road. All photographers recorded two 7 minute video sequences on a single day using two different cameras (and caps). 20 photographers also recorded another sequence a week later. The use of different cameras for different sequences came to ensure that motion rather than camera calibration is learned. No effort was made to ensure that the same shoes would be used on both days (and in fact several persons had changed shoes between sequences).

6.2. Evaluation Protocol

6.2.1 Photographer Identification

Photographer identification sets to recognize a photographer from a closed set of M candidates. For this task it is assumed that we have training data from all subjects.



Figure 13. The apparatus used to record the EVPR dataset.

We tested our method both on the FPSI the EVPR datasets. In the FPSI dataset we used for each individual the first 80% of sequences (taken in the morning) for training, and the last 20% sequences recorded in the afternoon for testing. This is done to reduce overfitting to a particular time or camera setup. Data were randomly sub-sampled to ensure equal number of examples for each photographer in both training and testing sets. The results are described in Sec. 4.

For the EVPR dataset we used sequences from Camera D1 for training. For testing we use both sequences from Camera D2 (taken on the same day) and Camera D3 (taken a week later, when available). The results on each camera are compared to analyze whether recognition performance degrades within a week.

6.2.2 Photographer Verification

Given a target photographer with a few minutes of training data, and negative training examples by other non-target photographers, we verify whether a probe test video sequence was recorded by the target photographer. Recognition on longer sequences is done by combining the predictions from subsequent short sequences. As the FPSI dataset contains only 6 photographers it was not suitable for this task (this was elaborated upon in Sec. 5) therefore only the EVPR dataset was used for evaluating performance on this task. For each of 32 photographers: i) photographer is designated target ii) we selected sequences of the target photographer and 15 non-target photographers (randomly selected) for training a binary classifier. All training sequences were 7 minutes (200 descriptors) long and were recorded by camera D1. iii) Another sequence recorded by the target photographer and the remaining 16 participants that were not used for training, were used to test the classifier. Test sequences were recorded by camera D2. iv) The ROC curve and EER were computed. Average EER and ROC for all photographers is finally obtained. As each sequence contained about 200 descriptors this formed a significant test set. Care was taken to ensure that all photographers (apart from the target) would appear in the training or test datasets but not in both. This was done to ensure we did not overfit to specific non-target photographers. We replicated positive training examples to ensure equal numbers of negative and

positive training and test data.

6.3. Implementation Details

Features: In all experiments the optical flow grid size used was 10×5 . In the CNN experiments, all optical flow values were divided by the square-root of their absolute value, this was found to help performance by decreasing the significance of extreme values. Feature vectors of length 60 frames at 15 fps (4s) were used. Feature vectors were extracted every 2s (with a 2s overlap).

Normalization: We followed the standard practice - For the LPC descriptor, all feature vectors were mean and variance normalized across the training set before being used by the SVM. For the CNN, feature vectors were mean subtracted before being input to the CNN.

Training: The SVM was trained using LIBSVM [4]. We used $\sigma = 1e - 4$ and $C = 1$ for LPC, $C = 10$ for the raw features. The CNN was trained by AdaGrad [5] with learning rate 0.01 on a GPU using the Caffe [9] package. The mini-batch size was 200.

7. Conclusion

A method to recognize the photographer of head-worn egocentric camera video has been presented. We show that photographer identity can be found from body motion information as expressed in camera motion when walking. Recognition was done with both physically motivated hand designed descriptors, and with a Convolutional Neural Network. Both methods gave good recognition performance. The CNN classifier was shown to generalize and improve on the LPC hand-designed descriptor.

The time-invariant CNN architecture presented here is quite general and can be used for other video classification tasks relying on coarse optical flow.

We have tested the effects of simple 2D video stabilization on classification accuracy, and found only slight degradation in performance. It is possible that more elaborate stabilization would have a greater effect.

The implication of our work is that photographers' head-worn egocentric videos give much information away. This information can be used benevolently (e.g. camera theft prevention, user analytics on video sharing websites) or maliciously. Care should therefore be taken when sharing such video.

Acknowledgment. This research was supported by Intel-ICRC and by the Israel Science Foundation.

References

- [1] Y. Bar, N. Levy, and L. Wolf. Classification of artistic styles using binarized features derived from a deep

- neural network. In *ECCV 2014 Workshops*, pages 71–84, 2014. 2
- [2] Y. Bengio. Learning deep architectures for AI. *Foundations and trends in Machine Learning*, 2(1):1–127, 2009. 2
- [3] J. P. Campbell Jr. Speaker recognition: a tutorial. *Proceedings of the IEEE*, 85(9):1437–1462, 1997. 1, 3
- [4] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM Trans. on Intelligent Systems and Technology*, 2011. 8
- [5] J. Duchi, E. Hazan, and Y. Singer. Adaptive subgradient methods for online learning and stochastic optimization. *JMLR*, 2011. 8
- [6] A. Fathi, J. K. Hodgins, and J. M. Rehg. Social interactions: A first-person perspective. In *CVPR*, 2012. 2, 5, 7
- [7] S. Furui. Cepstral analysis technique for automatic speaker verification. *ICASSP*, 1981. 2
- [8] J. Hays, A. Efros, et al. Im2gps: estimating geographic information from a single image. In *CVPR’08*, pages 1–8, 2008. 2
- [9] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. *arXiv:1408.5093*, 2014. 8
- [10] C. R. Johnson Jr, E. Hendriks, I. J. Berezchnoy, E. Brevdo, S. M. Hughes, I. Daubechies, J. Li, E. Postma, and J. Z. Wang. Image processing for artist identification. *IEEE Signal Processing Magazine*, 25(4):37–48, 2008. 2
- [11] K. M. Kitani, T. Okabe, Y. Sato, and A. Sugimoto. Fast unsupervised ego-action learning for first-person sports videos. In *CVPR*, 2011. 2
- [12] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *NIPS*, 2012. 2
- [13] S. Lee, H. Zhang, and D. J. Crandall. Predicting geo-informative attributes in large-scale image collections using convolutional neural networks. In *WACV’15*, pages 550–557, 2015. 2
- [14] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *CVPR*, 2013. 2
- [15] B. D. Lucas, T. Kanade, et al. An iterative image registration technique with an application to stereo vision. In *IJCAI*, volume 81, pages 674–679, 1981. 3
- [16] J. Mantyjarvi, M. Lindholm, E. Vildjiounaite, S.-M. Makela, and H. Ailisto. Identifying users of portable devices from gait pattern with accelerometers. In *ICASSP*, 2005. 1, 2
- [17] H. Murase and R. Sakai. Moving object recognition in eigenspace representation: gait analysis and lip reading. *Pattern Recognition Letters*, 17(2):155–162, 1996. 2
- [18] K. Nishino and S. K. Nayar. Corneal imaging system: Environment from eyes. *IJCV*, 70(1):23–40, 2006. 2
- [19] Y. Poley, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *CVPR*. 2, 7
- [20] Y. Poley, C. Arora, and S. Peleg. Head motion signatures from egocentric videos. In *ACCV*, 2014. 2, 7
- [21] D. A. Reynolds and R. C. Rose. Robust text-independent speaker identification using gaussian mixture speaker models. *IEEE Trans. on Speech and Audio Processing*, 3(1):72–83, 1995. 2
- [22] M. S. Ryoo and L. Matthies. First-person activity recognition: What are they doing to me? In *CVPR*, 2013. 2
- [23] K. Shiraga, N. T. Trung, I. Mitsugami, Y. Mukaigawa, and Y. Yagi. Gait-based person authentication by wearable cameras. In *INSS*, 2012. 2
- [24] C. Thomas and A. Kovashka. Who’s behind the camera? identifying the authorship of a photograph. *arXiv:1508.05038*, 2015. 2
- [25] R. Yonetani, K. M. Kitani, and Y. Sato. Ego-surfing first person videos. In *CVPR*, 2015. 2, 7